Improving Robustness using Generated Data

Year: 2021 | Citations: 340 | Authors: Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, D. A. Calian

Abstract

Recent work argues that robust training requires substantially larger datasets than those required for standard classification. On CIFAR-10 and CIFAR-100, this translates into a sizable robust-accuracy gap between models trained solely on data from the original training set and those trained with additional data extracted from the"80 Million Tiny Images"dataset (TI-80M). In this paper, we explore how generative models trained solely on the original training set can be leveraged to artificially increase the size of the original training set and improve adversarial robustness to ℓ_p norm-bounded perturbations. We identify the sufficient conditions under which incorporating additional generated data can improve robustness, and demonstrate that it is possible to significantly reduce the robust-accuracy gap to models trained with additional real data. Surprisingly, we even show that even the addition of non-realistic random data (generated by Gaussian sampling) can improve robustness. We evaluate our approach on CIFAR-10, CIFAR-100, SVHN and TinyImageNet against ℓ_∞ and ℓ_2 norm-bounded perturbations of size $\epsilon = 8/255$ and $\epsilon = 128/255$, respectively. We show large absolute improvements in robust accuracy compared to previous state-of-the-art methods. Against ℓ_∞ norm-bounded perturbations of size $\epsilon = 8/255$, our models achieve 66.10% and 33.49% robust accuracy on CIFAR-10 and CIFAR-100, respectively (improving upon the state-of-the-art by +8.96% and +3.29%). Against ℓ_2 norm-bounded perturbations of size $\epsilon = 128/255$, our model achieves 78.31% on CIFAR-10 (+3.81%). These results beat most prior works that use external data.